自有知识增强下的学术全文本关系抽取研究*

■ 卓可秋[1]　沈思[2]　王东波[1]

[1] 南京农业大学信息管理学院　南京 210095　[2] 南京理工大学经济管理学院　南京 210094

摘　要：[目的/意义]学术全文本下的关系抽取是学术全文本知识图谱构建的关键技术，所构建的学术知识图谱能够实现文献的结构化、知识化，提高研究人员检索文献、分析文献和把握科研动态的效率，以及通过图谱的认知推理，有助于隐式知识发现。[方法/过程]通过外部知识来增强关系抽取已在不少研究取得成果，但针对特定领域的关系抽取往往缺少可用的外部知识。研究发现，全文本中自有的高置信度的知识也可以用来辅助全文本关系抽取。受认知过程双系统理论（系统1为直觉认知，系统2为推理认知）启发，设计一个句子级模型来获取知识，并通过远程监督方式获取高置信度知识，然后将高置信度知识融入到全文本级深度学习模型最后分类的一层上。[结果/结论]在生物医学学术全文本数据集（CDR-revised）上，比当前最先进的模型在F1上提高11.13%。

关键词：学术全文本　关系抽取　自有知识增强　知识图谱

分类号：G250

DOI：10.13266/j.issn.0252-3116.2022.07.012

1 引言

关系抽取的目的是为非结构化文本中的实体分配语义关系。根据文本处理的范围，关系抽取可分为句子级关系抽取和全文本级关系抽取。句子级关系抽取的目的是获取句子中两个已知实体之间的关系。而全文本级关系抽取的目标是在包含多个句子的长文本中提取多个实体之间的关系。全文本级关系抽取样例如图1所示。关系抽取是知识图谱构建的关键技术。学术全文本的知识图谱构建自然离不开对学术全文本下的关系抽取研究。自Google于2012年提出知识图谱用于搜索引擎项目后，知识图谱就逐步替代语义网成为人工智能领域的一大研究热点。当前，知识图谱在语义搜索、智能问答、知识工程、数据挖掘和数字图书馆等领域有着广泛的应用。而将知识图谱与学术全文本进行结合，其中存在许多值得研究的、有价值的课题。

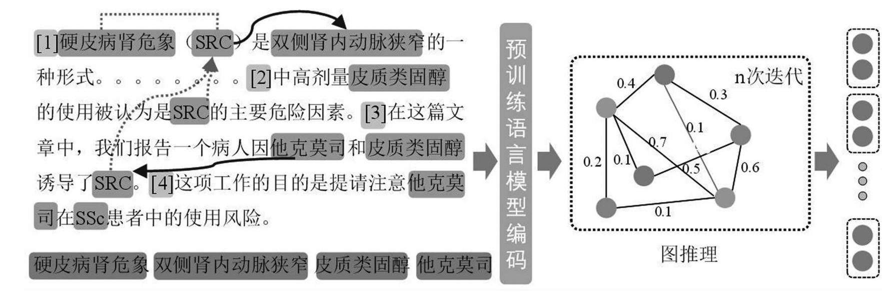


图1　全文本级关系抽取方法示例

学术全文本知识图谱的构建从是否围绕学术研究成果（也就是一篇论文）视角看可粗略地分为两方面：一是学术研究成果的宏观图谱构建，二是成果内所含学科知识的微观图谱构建。学术研究成果的宏观层面主要包括研究者的相关信息、研究成果发表的载体相关信息、研究方法、研究问题、研究结果、研究展望、引

* 本文系江苏省自然科学基金青年项目“基于深度学习的学术全文本时态语义知识标识及检索模型构建研究”（项目编号：BK20190450）和国家自然科学基金面上项目“基于深度学习的学术全文本知识图谱构建及检索研究”（项目编号：71974094）研究成果之一。

作者简介：卓可秋，博士研究生；沈思，副教授，博士生导师；王东波，教授，博士生导师，通信作者，E-mail：db.wang@njau.edu.cn。

收稿日期：2021-11-24　修回日期：2022-01-19　本文起止页码：120-131　本文责任编辑：徐健

文的研究问题、引文的研究结果、引文的研究方法等关系。学术研究成果内所含学科知识主要指领域知识。例如生物医药领域，在一篇研究论文[1]中指出他克莫司这种化学物质能诱导硬皮病肾危象，从该研究中可得出化学物质他克莫司与疾病硬皮病肾危象有一定关系。笔者所研究的关系抽取，是针对第二种情况，即研究从学术研究成果内抽取出含学科知识的关系。

在过去的几年中，人们已经提出许多方法来完成关系抽取任务，包括传统的依赖于人工特征工程方法[2]以及基于神经网络的模型方法[3-4]，基于神经网络的模型方法通过端到端的训练来提取特征并达到最先进的性能。这些基于神经网络的方法利用位置特征来获取实体信息。具体地，位置特征给出每个词与两个实体的相对距离作为模型的输入。最近的一些工作将预先训练的模型（如 BERT[5]）应用到关系抽取中。由于全文本级关系抽取中存在多个目标实体，因此一次性提供两个实体信息的实体标记方法不再适用，因为它们不能一次性提供所有实体的信息。

为了解决上述问题，大多数关于全文本级关系抽取的方法都是基于图模型进行适当地修正以达到较好效果[6]（见图 1）。具体来说，它们使用单词作为节点，句内和句间依赖（依赖结构、共指等）作为边。图模型提供了提取实体对特征的统一方法。后来的工作通过改进神经网络结构[7-8]或添加更多类型的边[9-10]来扩展图模型方法。但是全文本级中的实体关系抽取极其复杂，有些实体关系仅从全局加局部混合建模难以达到理想结果。如果有一些额外知识来辅助关系判断，将能得到更好的效果。

为此，笔者提出一种新颖的自有知识获取模型，以增强学术全文本关系抽取过程中的关系判断。具体来说：第一步是将处于同一个句子中的实体对进行关系判断。判断过程除了利用 SciBERT[11] 模型外，还增加了一个多视图的图模型，以增强关系的判别，此外，在得到句子中的实体对关系后将其放到搜索引擎中，采用远程监督的方式进一步确认，最终留下高置信度的实体对关系，作为自有的高置信度知识，简称自有知识。需要指出的是，本文的自有知识由于未经过人工校验，存在一定误差，因此不能称为真正意义上的知识，因为所谓知识是指清楚的、事实性的信息。第二步是对学术全文本进行实体提及编码、实体提及与句子的推理建模、自适应阈值选取以及自有知识增强抽取实体对关系等一系列操作，完成全文本级的关系抽取。简言之，第一步的作用是获取自有知识，第二步的作用是先获得初步的全文本级的实体对关系后利用第一步的自有知识来提升最终结果的准确性。

需要说明的是，以上的方法结合了认知科学中的双系统理论[12]。在人脑的认知系统中存在两个系统：系统－1 和系统－2。系统－1 是一个直觉系统，它可以通过人对相关信息的一个直觉匹配寻找答案，非常快速、简单。这个过程对应到本文中从句子抽取关系知识模块，因为从全文本上看，句子层面的关系更加直观。而系统－2 是一个分析系统，它通过一定的推理、逻辑找到答案，这个过程对应到本文的全文本关系抽取模块，因为该模块需要涉及各种语言逻辑推理。本文的主要贡献如下所示：

（1）提出了一个从学术全文本中获取自有知识来增强学术全文本的关系抽取。这是第一项从自有知识的角度，而不是外部的知识（如知识图谱）来增强关系抽取的工作，同时也是对双系统认知理论的一种验证。

（2）引入了多视图的图模型、多路径推理网络和自适应阈值选取等先进的技术，确保了学术全文本关系抽取的精度。与已有常见深度学习关系抽取方法相比抽取效果提高显著。

（3）在生物医学学术文本 CDR-revised 和 GDA 数据集上的实验结果表明所提出方法的有效性，尤其是优于最近的一些基线模型。促进学术全文本知识图谱构建中关系抽取关键技术的进一步完善，加快学术全文本的检索以及知识构建的落地。

2 相关工作

2.1 深度学习下关系抽取技术演化

采用文献内容分析法，笔者总结出深度学习下关系抽取的技术演进。如图 2 所示，关系抽取技术的演进主要体现在单词特征、外部知识、深度模型、训练数据集规模、抽取效果和研究领域等方面。在单词特征方面，先是引入了单词、词性、语法关系、Wordnet 超词等特征，之后加入实体类型标识和单词位置等特征。实体类型的引入能够缩小关系的类别范围，单词位置的引入能够体现词与词之间的上下文语义信息。在外部知识方面，一开始常用的方法是远程监督和迁移学习，近两年更倾向知识图谱的融合研究。事实上远程监督、迁移学习和知识图谱融合都能在一定程度上提高关系抽取的精度，但由于知识图谱能提供更有效的辅助信息，再加上各领域知识图谱的逐步完善，因而研究者们更青睐知识图谱融合方式作为外部知识来提高精度。

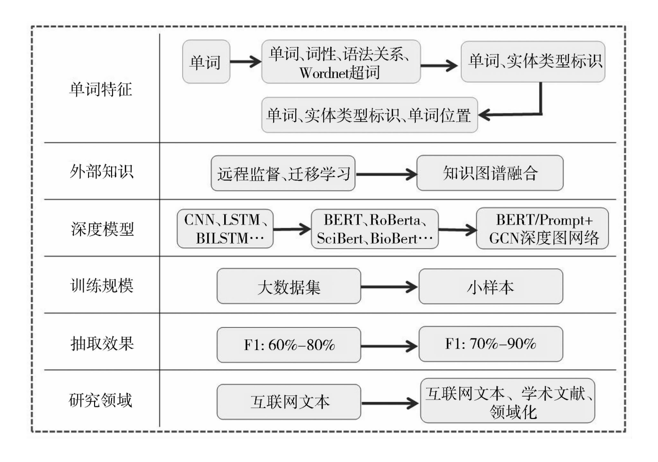


图 2　深度学习下关系抽取技术演进

在深度模型方面，受到通用深度学习模型的逐步改进，关系抽取的研究也不断地引入最新的、高效的通用模型以增强抽取效果。在训练规模方面，同样受到通用预训练语言模型的影响，关系抽取的研究从原来较大的训练数据集转到较小数据集上，以适应语料标注成本大的现实问题。在抽取效果方面，经过 5 – 6 年的发展，F1 得分提高约 10%，效果显著。在研究领域方面，人们先是采用了较易获取的互联网文本（如维基百科）作为研究对象，后来随着研究阵营的壮大，关系抽取研究在各学科子领域逐步开花，构建出各领域特有的知识图谱。

2.2　关系抽取相关研究

由于全文本级的关系抽取相较句子级的关系抽取复杂很多，所以一开始人们主要集中在句子级的关系抽取研究上[13 – 15]。句子级关系抽取的目的是检测句子中实体之间的关系。现有的句子级关系抽取模型可以分为两类：基于序列的和基于依赖关系的。基于序列的模型仅对单词序列进行操作，可以是单向序列也可以是双向序列[16 – 17]。基于序列的模型方法实现较简单，但易受上下文其他字、词影响，难以有效准确捕获目标实体对的语义关系。基于依赖性的模型将依赖树合并到模型中，从设计思路上看这类模型能够避免基于序列的模型易受上下文其他字、词的干扰，但实现困难，原因主要有两方面：①依赖树主要依赖句法分析生成，容易导致误差累加；②将依赖树模型融入深度学习模型，这种融合技术目前还尚未成熟。L. B. Soares 等研究发现通过实体标记（entity marker）方式能有效提高关系抽取的准确性[18]。这种实体标记方式在后续研究中被普遍应用。

然而在通常的文章行文中，很难在一句话中把关系描述清楚。特别在描述中含有多个实体和多个关系时，一般需要通过多句话来表达。这就要求有能力在更长的上下文句子间提取关系。最近的工作开始探索全文本级的关系抽取。Y. Yao 等利用 Wikipedia 和 Wikidata 公开的一个大规模通用数据集 DocRED，使得全文本级关系抽取有了很大的进展[19]。关于全文本级关系抽取的大多数方法都是基于图神经网络来捕获句子级间的语义信息[7,20 – 22]。F. Christopoulou 等通过共现和启发式规则构造了包含不同粒度（句子、提及、实体）的图，在没有外部工具的情况下对图进行建模[23]；S. Zeng 等构建不同粒度的双图来捕获文档感知特征和实体之间的交互[24]；Z. Guo 等提出一种细化机制来实现对整个文档的多跳信息进行聚合，所提的 LSR 模型在全文本级关系抽取上实现了较好的性能[25]。图神经网络利用图结构表达节点间依赖关系的先天优势，能够一定程度上解决序列模型受上下文其他字、词间的干扰的问题，但鉴于实际计算硬件影响，图神经网络往往需要将完整的一张大图切分成若干小图，适应小批量端到端的运算机制，这种大图中找小图的相关实现技术还有待进一步研究。

W. Xu 等设计了一个判别式推理网络，根据所构造的图和每个实体对的上下文向量，估计不同推理路径的关系概率分布，从而识别实体对之间的关系[22]。ATLOP 模型是已知在当前开源全文本级语料 DocRED 上效果最好的模型[26]。该模型提出了一种自适应阈值技术，用一个可学习的阈值类代替全局阈值。这种技术消除了阈值调整的需要，并且使得阈值可以根据不同的实体对进行调整，从而获得更好的结果。

利用已有的知识图谱来指导关系抽取是关系抽取的另一个发展方向[27 – 29]。知识图谱中蕴含的大量实体关系信息，可以有效弥补关系抽取训练过程中的数据不充分问题。融合知识图谱的关系抽取方法，主要有以下几类：①从模型特征的角度融合，将实体类型信息加入到注意力机制中，让关系抽取模型能够更有效地捕捉文本语义特征，从而提升关系分类效果；②从监督训练的角度融合，对知识图谱进行预训练，利用知识图谱的嵌入表示对关系抽取模型进行监督，有效降低了目前关系抽取训练集中的噪音信号；③从类别推理的角度融合，进行了零样本学习的探索，通过知识图谱的层次结构，捕捉关系和关系之间的相关性，让某些训练数据极少的关系也能被充分识别。

Q. Chen 等运用同义词、反义词、上下义词和共下义词的基本知识，帮助建立句子对之间的软对齐神经网络模型[30]。然而，它只能处理固定数量的知识类

型,而且在训练前需要预先为关系赋值,这限制了它在实践中的应用。Z. Wang 等提出了一种知识图增强的自然语言推理(KGNLI)模型[31],KGNLI 首先从给定的句子对中提取诸如主语、谓词和对象之类的实体,然后基于包含这些实体作为节点的知识图学习知识关系表示。此外,KGNLI 还通过双向长短时记忆(BiLSTM)网络学习给定句子之间的语义关系表示。最后,KGNLI 将这两种表示结合起来,并将其输入多层感知器以确定关系的标签。但句子对间决定它们之间关系的关键字很难被找到,KGNLI 也不例外。M. E. Peters 等提出一个知识注意力和上下文重构(KAR)组件,对 BERT 模型进行改造提升 BERT 模型的能力,从而提升关系抽取的能力[32]。但 KAR 仅将知识图谱上的词嵌入与序列中的词嵌入进行融合,无法充分发挥知识图谱上实体对的关系知识,进而影响待判别的实体对所在上下文的语义表达能力。

3 研究方法

3.1 自有知识增强下的学术全文本关系抽取框架

学术文本与非学术文本相比,从行文上来说更加严谨,立论的论点和论据更明确,逻辑更清晰,由此带来句子更复杂,上下文的逻辑推理关系也较多。为此,笔者设计出一个自有知识增强下的学术全文本关系抽取模型来提高学术全文本中关系抽取的准确性。如图 3 所示,自有知识增强下的学术全文本关系抽取框架分为左右两大部分。图的左半部分用于获得句子级的自有知识,图的右半部分用于获得全文本级的实体对关系,且左半部分的自有知识用来指导右半部分的全文本级关系抽取。

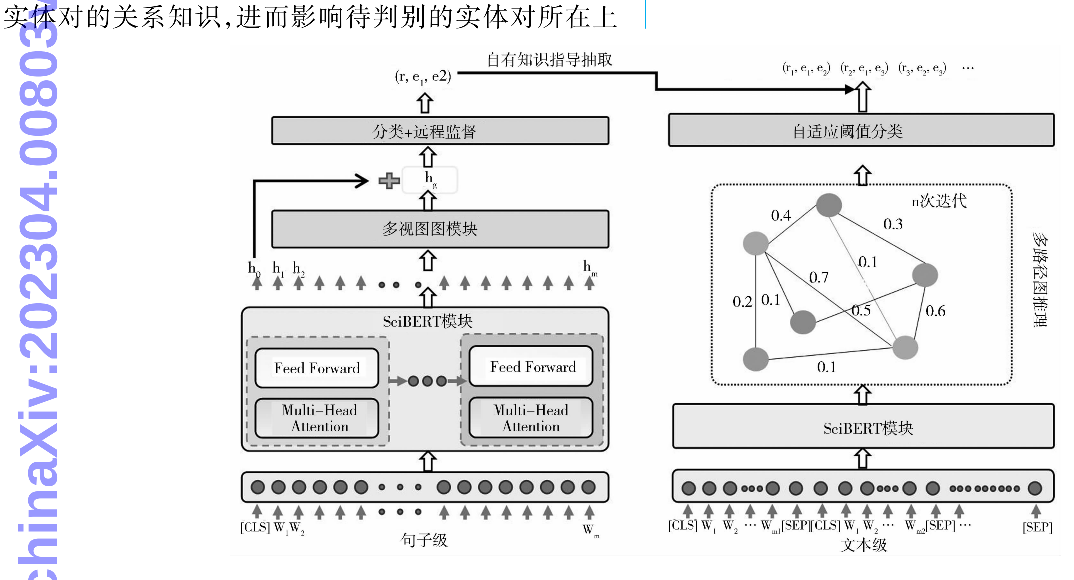


图 3　自有知识增强学术全文本关系抽取框架

在学术全文本关系抽取时,不仅要考虑句内的语义逻辑还要考虑句间语义逻辑,而且比非学术文本更加注重逻辑推理这部分。基于此,笔者先从推理入手来解决学术全文本的关系抽取问题(对应图 3 右半部分)。下面详细介绍该方法。

3.2 基于推理的学术全文本关系抽取

学术全文本的关系抽取涉及多种推理,主要包括模式匹配、逻辑推理、共指推理以及常识推理[33]。当前大部分的相关研究都仅采用一个图模型试图将实体对经过多跳图卷积后得到相应的低维分布式表达,并以此计算两者之间的关系,完成各种类型的推理。这种方式虽然能一定层度上达到较好的推理效果,但忽略了一个技巧就是不同推理形式需要不同的建模策略。笔者参考 W. Xu 等的做法,将推理分为句内推理、逻辑推理和共指推理并对这 3 种推理建立不同的推理路径并建模[22]。

通常一个实体对包含多个实体提及,也就是说一个实体对之间的关系,通过以上的 3 种推理方式能够产生好多种关系。其中逻辑推理部分由于实体提及 m_k 可能存在多个,此时这部分的概率也会有多个。本文的策略是,留下一个概率最大的实体提及对的关系作为实体对可能的关系。只有当关系的概率大于某一个阈值时才输出该关系,否则输出无的关系。下面介绍如何获得该阈值。

一般来说,会对一个全局的概率阈值与概率 Pr 做比较,来最终判断实体对所属的类别。全局概率阈值的获取一般通过多次试验计算验证集的 F1 指标,当 F1 指标最大时,从而获得一个全局概率阈值。这种方

法存在两个弊端:①不一定适合所有类别,如类别 1 需要的概率阈值是 0.5,而类别 2 需要的概率阈值可能是 0.4;②需要多次运行验证集获取全局概率阈值,时间复杂度较高。

参照 W. Zhou 等的做法,笔者用一个可学习的阈值类别代替全局阈值[27]。具体做法是在模型训练中,引入虚拟阈值类别,将正例类别和负例类别分隔开来。要求所有正例类别的概率都比虚拟阈值类别的概率高,同时所有负例类别的概率都比虚拟阈值类别概率低。在模型推断时,返回概率高于虚拟阈值类别的类别作为预测类别标签,或者如果没有高于虚拟阈值类别时,就返回无这一类别标签。这种技术使得阈值可以根据不同的实体对的关系类别进行调整,从而获得更好的结果,同时降低了多次运行验证集获取阈值的时间复杂度。

笔者研究发现仅靠上述的推理方式识别学术文本中的关系,虽然能一定程度上解决需要句间推理与句内推理的问题,但由于实体间的超长依赖,导致推理过程易受上下文其他字、词、句的干扰。为了解决该问题,笔者提出再利用一个句内推理能力强的独立模型来进一步提升学术全文本的关系抽取。下面详细介绍这个独立模型,即自有知识获取模型(对应图 3 左半部分)。

3.3 自有知识获取模型

受认知科学双系统理论启发,笔者尝试先从句子中获取简单的、明确的实体关系,然后再以此关系辅助全文本关系抽取中复杂的推理。

句子中的关系通常比较明确,但是遇到多实体时,关系的自动识别效果就会大打折扣。如何保证所获取的关系具有较高可信度是值得研究的课题。一种做法就是只从少量实体的句子中训练模型并预测含有少量实体的新句子,以此获得新句子中的关系。一般来说这种方法所获得的句子关系准确率较高,但这种会漏掉很多关系知识。第二种做法是在第一种的基础上,再尝试从多实体的句子中发现关系,但这种方式具有较大难度,需要精心设计抽取的方法。笔者选取第二种方法来获取高置信度的实体对关系。从图 3(左边)看出,主要分为 3 个模块:BERT 模块、多视图的图模块和远程监督模块,以尽可能获得高准确性的实体对关系。从句子中所获取的实体对关系,就称为自有知识,能够用于后续的全文本关系抽取。下面详细介绍 3 个模块内容。

3.3.1 BERT 模块

笔者使用 SciBERT[34] 作为编码器来抽取句子的语义信息。SciBERT 采用学术语料库训练而成,它与普通的 BERT 模型在词汇上仅有 42% 的重叠,说明学术领域文本和一般领域文本之间常用词差异显著。SciBERT 已被证明在学术文本上的各种语言任务上其性能优于普通 BERT 模型。将句子中的单词以 token 的形式输入 SciBERT,经过编码,输出 $H = \{h_0, h_1, h_2, ..., h_{m-1}, h_m\}$。通常 h_0 会被用来代表句子,并作为类别标签的判断。笔者参照 F. Xue 等的方法,在 SciBERT 编码器输出的 H 上增加一层图模块,然后将学习到的图与 h_0 结合作为分类器的输入[35]。下面介绍所增加的一层图模块,即多视图的图模块。

3.3.2 多视图的图模块

(1)高斯图与卷积计算。BERT 模块输出的 $H = \{h_0, h_1, .. h_m\}$,其中 h_0 是由句首的[CLS]得来,无需传入图模块。将剩余的编码表达输入图模块,并将它们标记为 $V^0 = \{v_1^0, v_2^0, \cdots, v_m^0\}$。对 $v_i^0 (i = 1 \cdots m)$ 生成 N 个高斯分布:$\{N_1^0, N_2^0, \cdots, N_N^0\}$,其中高斯分布的期望和方差都是通过可训练的神经网络获得。之所以采用多视图方式生成高斯分布,主要有 2 个原因:①能够尽可能地获得 token 的各种含义;②在不知道 token 的先验分布时,选择高斯分布是一种比较保险的决策,因为中心极限定理说明很多独立随机变量的和近似服从高斯分布,很多真实分布本身就很接近高斯分布。

(2)动态时间池化与分类器模块。在每一层卷积之后都会接一个动态时间池化(DTWPool[35])。对于图的第 n 个视图,先计算各节点的注意力,然后采用 SAGPool[36] 的方法对节点进行筛选,所剩余的节点集合就是原有节点集合的子集。经过 L 层的池化从而得到 L 个图$\{G_1, G_2, \cdots, G_L\}$,其中每个图内的节点都是 N 个视图的并集。由于每个句子的长度不一致,使得图中含有效信息节点的个数不一致,这就需要有一种池化的机制将重要的节点信息保留下来。解决办法就是引入一个支持节点个数不一致的损失函数,使得 G_1 和 G_L 的差异最小化。这种方式能够最大层度捕获更多的局部信息。最终的结果图由各个层次的图归并而成。由于各层上图的节点个数不同,所以只选择与图 G_L 含相同节点或是 G_L 子集的图。

对于动态时间池化最终输出的图,再经过一层最大池化,便得到图的向量表示。该图的向量表示能够辅助 BERT 模块编码所得到的表示[CLS]的 h_0,让关系分类精度更准。将 h_0 和图向量拼接,再经过一个 softmax 层,就能得到句子中一个实体对的关系类别标签。当句子含多个实体对时,重复以上计算,每次只判

断一对实体的关系。

3.3.3 远程监督模块

采用 BERT 模型可以学习到基础的语言知识,改善抽取的结果,但依旧摆脱不了需要依赖于打标的数据。针对上述问题,一种称为远程监督的方法应用而生[37－40]。远程监督也称为弱监督。关系抽取中的远程监督方法主要是利用知识图谱中的实体对和关系,到各种可获得的文本中比对,当实体对同时出现在文本,就将该文本当作是包含该关系。笔者也借用这种思想,将上文所获得的实体对放到搜索引擎中查找,当搜索结果的前 topn 条中的其中 μ 条同时包含实体对,就将实体对所对应的关系认为是正确的关系,否则就丢弃该实体对的关系。实验中,topn 取 10,μ 取 3。之所以 topn 取 10 是因为搜索引擎结果首页通常为 10 条记录,首页的信息已足够,不需要再用到其他页内容。至于 μ 取 3 请见下文的 4.6 小节。如图 4 所示,实体对是运动障碍和左旋多巴,它们在百度的搜索结果中同时出现,且符合筛选要求,此时就将该实体对保留,并用它们的关系作为知识来增强后续的全文本关系抽取。



图 4 远程监督模块样例

3.3.4 自有知识增强

当某个实体对在全文本关系抽取的结果中为无关系时,判断自有知识中所确定的关系,如果自有知识中存在关系,那么就把该实体对在全文本关系抽取的结果改为自有知识中的关系。采用此方法的原因主要有:首先,句内推理会受到其他句的一些上下文信息的干扰,所以需要一个单独的模型来获得句内的自有知识。其次,如果把所获取到的自有知识融入到全文本关系抽取模型内部,用于指导其他实体对的关系抽取,这对模型构造来说相当困难。再次,如果通过判断全文本关系抽取中实体对的关系概率小于一定阈值时,就将自有知识替换实体对的关系。这种情况带来一个问题就是阈值难以设定。最后,笔者直接采用当实体对在全文本关系抽取的结果中为无关系时替换成自有知识中的关系,这种方法虽然简单,但经过实验验证,确实有效。如图 3(右边)所示,全文本级 token 经过上文所述的基于推理的学术全文本关系抽取步骤后,经过自有知识增强模块来最终判定实体对的关系类别。

当实体对在全文本关系抽取的结果中为无关系时被替换成自有知识中的关系的合理性的原因有以下两点:①与句子级关系抽取相比,全文本下的关系抽取更具挑战性。因为后者不仅需要考虑句内实体间的逻辑关系,还要考虑跨句间实体对逻辑关系,涉及实体对的长依赖问题,挑战性更大[19,41]。②从现有公开数据集看,句子级的实验结果明显比全文本级(也称为文档级)高 10% －20%[42],这一点也一定程度上说明句子级的关系结果的可信度高于全文本级的关系结果。因此,利用自有知识模块中抽取的简单的实体关系可以部分替代全文本级模块中提取出的结果是有其合理性。

4 实验与分析

4.1 学术全文本关系抽取数据集与模型参数设置

全文本关系抽取相关的公开数据集主要包括 DocRED[43]、CDR[44] 和 GDA[45]。所有这些都涉及到跨多个句子的多个实体的关系推理,具有极大挑战性。DocRED 是由 Wikipedia 和 Wikidata 构建的大规模数据集。CDR 是使用 PubMed 构建的生物医学数据集,涵盖化学物质与疾病的二元关系,对生物医学研究具有重要意义。GDA 数据集也是一个二元关系分类任务,用于识别基因和疾病的相互作用,它在 MEDLINE 上采用远程监督方式构建而成,数据集的质量与 CDR 相比较差。笔者研究的是学术全文本下的关系抽取,所以在以上 3 种公开数据集中选择了 CDR 和 GDA 两个数

据集进行实验。

表1为数据集CDR和GDA的数据量特征统计。CDR数据集总共包含1 500篇文本，平均分成3份，包括训练集、验证集和测试集，所有数据都由人工标注得来。GDA数据集总共包含30 192篇文本。与CDR相比，每篇文本平均含实体数少2，每篇文本平均含实体提及数少0.7，而每个句子平均含实体提及数基本一致。需要说明的是，笔者在研究中发现CDR测试数据集中的很多实体对关系没有标注出来，被认为是无关系，笔者针对这些无关系的实体对，重新进行人工校验，修正了161个实体对关系，这一部分测试集称之为CDR-revised。

表1 数据源CDR和GDA的数据量特征统计

统计	CDR	GDA
训练集数量	500	23 353
验证集数量	500	5 839
测试集数量	500	1 000
关系数	2	2
每篇文本平均含实体数	6.8	4.8
每篇文本平均含实体提及数	19.2	18.5
每个句子平均含实体提及数	2.48	2.28

笔者采用Apex的混合精度训练方法[46]进行模型训练。在表2中列出了所涉及的一些超参数。所有超参数都在开发集上进行调整。

表2 模型参数设置

统计	CDR	CDR-revised	GDA
批量大小(batch size)	4	4	4
学习率(learning rate)	5e-5	5e-5	2e-5
轮次(epoch)	{10,20,30,40}	{10,20,30,40}	{4,6,8,10,15,20}
词嵌套维度(word embedding size)	100	100	100
实体类别嵌套维度(entity type embedding size)	20	20	20
共指嵌套维度(coreference embedding size)	20	20	20
图卷积网络层数	2	2	2
权重消退系数	0.0001	0.0001	0.0001
优化器	AdamW	AdamW	AdamW

4.2 与其他模型对比实验

笔者对比了自有增强模型ESOKRE与其他类似的研究成果，包括BRAN[47]、EoG[48]、LSR[9]、DHG[49]、GLRE[50]、SciBERT base[51]和ATLOP-SciBERTbase[27]。从表3(5次试验，取F1得分最高的一次)可见，本文的自有知识增强模型ESOKRE与ATLOP-SciBERTbase相比，在CDR-revised和GDA数据集上，F1得分分别提高11.13%和0.35%。在CDR数据集上F1偏低，是因为该数据集标注结果不全，当自有增强模型判断出的实际为正确的关系，由于在该数据集上没有标注而被误认为是无关系。在GDA数据集上本文的模型方法提高不明显，是因为GDA数据集中基因与疾病的关系在行文中相对明确，应用较为简单的推理既能达到较好效果，所以这种情况下本文的自有知识增强能力就很难得以体现。从表4可知，在CDR和CDR-revised数据集上自有知识增强修正的关系数占比都为7.07%。这两个数据集修正的关系数相同但F1结果不同，是因为CDR-revised是在CDR基础上修正过的。GDA数据集上自有知识增强修正的关系数占比为0.40%，修正数量有限，从而影响最终的F1得分提高。

表3 数据集CDR、CDR-revised和GDA上各模型F1结果对比

模型	CDR	CDR-revised	GDA
BRAN	62.10	-	-
EoG	63.60	-	81.50
LSR	64.80	-	82.20
DHG	65.90	-	83.10
GLRE	68.50	-	-
SciBERTbase	65.10	-	82.50
ATLOP-SciBERTbase	**69.40**	64.70	83.44
ESOKRE	67.90	**75.83**	**83.83**

表4 自有知识增强前后关系数占比统计

指标	CDR	CDR-revised	GDA
关系数	5 204	5 204	5 222
增强后修正的关系数	368	368	21
增强后修正的关系数占比	7.07%	7.07%	0.40%

4.3 消融实验

笔者进行了模块消融实验，以验证所提方法不同组成部分的有效性。从表5的CDR-revised数据集中，可观察到不管缺失哪个模块性能都有所下降。ESOKRE是指自动获取自有知识并通过远程监督方式增强全文本关系抽取。“ESOKRE - 自有知识增强”意味着仅采用全文推理，而不是融合了自有知识。“ESOKRE - 自有知识中SciBERT”是指自有增强模型中BERT模块采用RoBERTa而不是SciBERT。“ESOKRE - 自有知识中图模块”是指自有增强模型中少了多视图的图模块。“ESOKRE - 远程监督”是指自有增强模型中少了远程监督模块。从CDR-revised数据集结果可见，自有知识增强模块和BERT模块中使用SciB-

ERT 对模型性能贡献最大。若把它们从整个模型方法中删除时，F1 得分分别下降 6.72% 和 4.32%。这表明笔者提出的自有知识增强模块辅助全文本关系逻辑推理是有效的。此外，自有知识增强模块中如果去掉远程监督子模块，F1 得分下降了 1.91%，这说明远程监督子模块能一定程度上提高自有知识获取的精度。

表 5　数据集 CDR 和 CDR-revised 上的知识增强模型 ESOKRE 消融实验

模型	CDR			CDR-revised		
	Precsion	Recall	F1-score	Precsion	Recall	F1-score
ESOKRE	68.17	69.13	67.90	70.04	82.68	75.83
ESOKRE － 自有知识增强	58.95	80.07	68.64	68.17	70.08	69.11
ESOKRE － 自有知识中 SciBERT	65.37	63.23	64.28	65.7	78.45	71.51
ESOKRE － 自有知识中图模块	66.05	67.1	66.57	68.56	81.32	74.39
ESOKRE － 远程监督	63.22	68.14	65.58	67.34	81.92	73.92

需要说明的是，表 5 中的 CDR 数据集增加了自有知识增强模块后，反而 F1 得分下降，其原因在 4.1 节也有提过，即 CDR 测试数据集中的很多实体对关系没有标注出来，被认为是无关系，从而影响了实验结果。由于笔者所提方法在 GDA 数据集提高不明显，所以本小节及后续小节不再实验该数据集。提高不明显的原因已在上一小节提到。

4.4　训练量大小对实验结果的影响分析

在学术全文本知识图谱构建领域，对于训练数据的标注十分耗时耗力。因此需要评估在有限的标注数据上，学术关系抽取模型性能的提升情况。笔者对比了 ATLOP-SciBERTbase 和 ESOKRE 两种模型方法随训练文本量的变化情况，实验结果如表 6 所示。两种模型方法在文本训练量从 10 篇逐渐调整至 500 篇时，F1 得分逐渐增长，采用 ATLOP-SciBERTbase 模型方法时，F1 得分从 13.66 增长到 63.87。采用笔者所提出的 ESOKRE 模型方法时，F1 的增长幅度和速率与 ATLOP-SciBERTbase 模型方法类似，具体地，F1 得分从 15.83 增长到 75.83。从 F1 数据的变化情况可见，训练数据量对学术文本的关系抽取影响较大，但也不是说只要一直增大训练数据量就能达到相应的准确性，因为训练量与准确性的关系不是线性递增的，这个结论对实践具有很大的指导意义，即在某领域的学术关系抽取工作中，需要评估人工标注的具体量。这个具体量的确定方法就是通过迭代的方式，逐步标注，训练模型，并评估效果，当评估值达到增长曲线上的拐点时，基本就能凭此确认待标注的最终数据量。

表 6　训练量大小对实验结果的影响对比

训练文本量/篇	ATLOP-SciBERTbase			ESOKRE		
	Precision	Recall	F1-score	Precision	Recall	F1-score
10	47.80	7.97	13.66	49.29	9.43	15.83
50	61.00	45.53	52.14	61.20	47.15	53.26
100	65.75	48.86	56.06	66.70	49.89	57.08
150	66.28	50.98	57.63	66.29	53.42	59.17
200	64.86	56.42	60.35	66.27	77.80	71.58
250	66.22	55.93	60.64	66.71	79.35	72.48
300	66.80	56.26	61.08	66.76	79.67	72.65
350	66.08	58.29	61.94	66.84	81.30	73.37
400	67.50	57.40	62.04	67.17	80.33	73.16
450	68.05	60.08	63.82	68.18	82.93	74.83
500	68.17	60.08	**63.87**	70.04	82.68	**75.83**

4.5　自有知识关系替代方式分析

上文研究方法章节提到使用自有知识关系替代候选的全文本关系的方式是通过实验验证获得，下面介绍该实验验证过程。笔者将自有知识关系替代候选的全文本关系的方式分为“无关系时替换”“有关系时替换”和“直接替换”3 种，分别表示当候选的全文本实体对关系的结果为无关系且自有知识有关系时进行替换、当候选的全文本实体对关系的结果为有关系且自有知识无关系时进行替换、以及直接用自有知识的关系替换候选的全文本实体对关系的结果。从表 7 可

见，在 CDR 和 CDR-revised 数据集上，无关系时替换优于直接替换，直接替换优于有关系时替换。之所以产生这种结果差异，笔者认为在实体对长依赖的逻辑推理中得到的无关系结果能够通过自有知识进行补充，而自有知识中没有的关系贸然替换通过长依赖推理出的关系本身就不合理，因为有些关系不会体现在句内，而是句间。

表 7　自有知识关系替代方式对比

关系替代方式	CDR			CDR-revised		
	Precision	Recall	F1-score	Precision	Recall	F1-score
无关系时替换	58.95	80.07	**67.9**	70.04	82.68	**75.83**
有关系时替换	73.59	51.35	60.49	73.59	44.63	55.56
直接替换	5.978	62.3	61.01	74.23	67.23	70.56

4.6　远程监督模块中参数 μ 取值对关系抽取结果的影响分析

上文提到在远程监督模块中当搜索结果的前 topn 条中的其中 μ 条同时包含实体对，就将实体对所对应的关系认为是正确的关系，否则就丢弃该实体对的关系。在实验中，topn 取 10，μ 取 3。μ 取 3 是因为取该值能获得最佳的 F1 得分。图 5 是在 CDR-revised 数据集上所做的远程监督模块中不同 μ 取值对关系抽取结果的影响对比，从图 5 中可见，μ 在 3 位置最终的关系抽取 F1 得分最高，为 75.83%，而取值小于 3 或者大于 3 时，关系抽取的 F1 得分都有所下降。

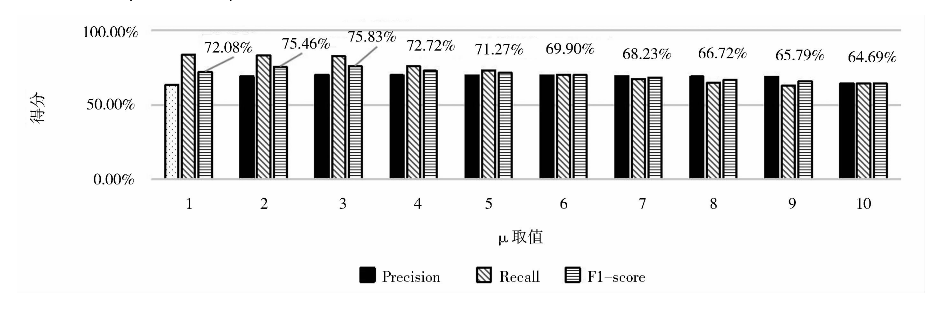


图 5　远程监督模块中参数 μ 不同取值的结果对比

4.7　案例研究

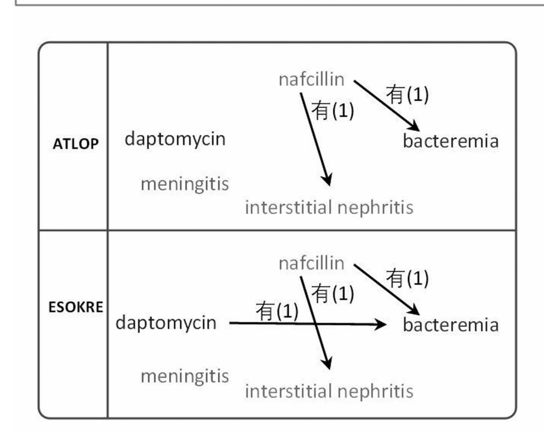
笔者通过案例研究以进一步说明所提模型 ESOKRE 与基线模型相比的有效性。如图 6 所示，ATLOP（ATLOP-SciBERTbase 的简称）和 ESOKRE 都可以成功提取实体对“nafcillin”和“interstitial nephritis”以及实体对“nafcillin”和“bacteremia”之间的“药物—疾病”关系。然而，只有本文的模型 ESOKRE 能够提取出实体对“daptomycin”和“bacteremia”之间的“药物—疾病”关系。从本案例可见，虽然基线模型在 CDR 全文本的实体关系矩阵上具有一定的推理能力，但针对个别情形就存在不足。例如本案例的“daptomycin”和“bacteremia”关系，从第［3］句中可推断出“daptomycin”和“bacteremia”含有“药物—疾病”关系，第［1］句虽然含有两者的实体提及，但难以明确判断二者的关系，此时通过全文本的推理，［1］句中的上下文就会对两者关系推断产生干扰。而本文的模型 ESOKRE 是先提取自有知识，通过该模块运算，就已经能将“daptomycin”和“bacteremia”的关系明确下来。

再如表 8 的 3 个对比案例，所列举的都是 ATLOP 模型失效而本文所提 ESOKRE 模型有效的例子。从表 8 中的第 1 个案例可以看出，其中实体“vancomycin”和“nephrotoxicity”在第［1］和［12］句中就能明确得出化学物质诱导疾病的关系。同样地，从表 8 中第 2 个案例可以看出，其中实体“acute renal failure”和“chinese herbal”在第［4］句中就能明确得出化学物质诱导疾病的关系。在表 8 中的第 3 个案例中，也是类似在第［4］句中即能得出实体“cerebral vasospasm”和“cytarabine”的关系。综述所述，可分析得出 ALTOP 模型之所以失效，是因为受上下文其他句子、实体的影响，按目前的长文本语义分析技术，仅靠单模型很难有效解决。

5　结语

笔者提出了一种利用学术全文本的自有知识来增强全文本的关系抽取的有效方法。该方法从句子中获得实体对关系后将其放到搜索引擎中，采用远程监督的方式进一步确认，最终留下高置信度的实体对关系，

[1] To report a case of methicillin-sensitive Staphylococcus aureus (MSSA) bacteremia with suspected MSSA meningitis treated with high-dose daptomycin assessed with concurrent serum and cerebrospinal fluid (CSF) concentrations.
[2] DISCUSSION: Daptomycin was initiated in our patient secondary to possible nafcillin-induced acute interstitial nephritis and relapsing bacteremia .
[3] CONCLUSIONS: High-dose daptomycin may be an alternative option for MSSA bacteremia with or without a CNS source in patients who have failed or cannot tolerate standard therapy.



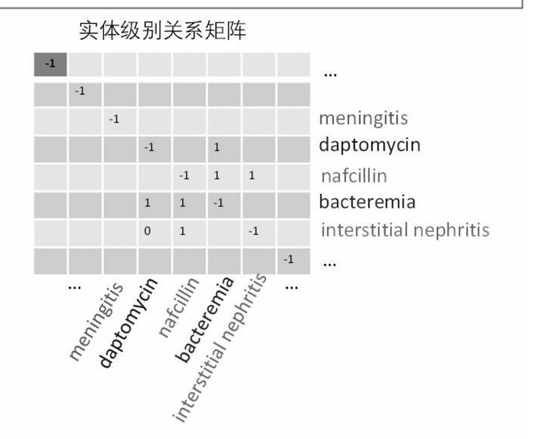


图 6　本文所提模型 ESOKRE 和基线模型的对比案例 1

注:图中特殊的数字表示关系类别 ID,左图中实体间未连线表示两者无关系

表 8　模型 ESOKRE 和基线模型的对比案例 2－4

序号	案例	实体 1	实体 2	ATLOP-SciBERTbase	ESOKRE
1	[1]In vivo evidences suggesting the role of oxidative stress in pathogenesis of **vancomycin**-induced nephrotoxicity: protection by erdosteine. ...[2]The aims of this study were to examine **vancomycin** (**VCM**)-induced oxidative stress that promotes ...[12]... important role in the **VCM**-induced nephrotoxicity and the modulation of oxidative stress with...	**vancomycin**	**nephrotoxicity**	无关系	化学物质诱导疾病关系
2	[1]**Acute renal failure** associated with prolonged intake of slimming pills containing anthraquinones. [2]**Chinese herbal** medicine preparations are widely available and often regarded by the public...[4]We report a 23-year-old woman who developed **acute renal failure** following prolonged use of a proprietary **Chinese herbal** slimming pill that contained anthraquinone derivatives, extracted from Rhizoma Rhei (rhubarb)...	**acute renal failure**	**chinese herbal**	无关系	化学物质诱导疾病关系
3	[1]Acute encephalopathy and **cerebral vasospasm** after multiagent chemotherapy including PEG-asparaginase...[4]... and right-sided weakness with diffuse **cerebral vasospasm** on magnetic resonance angiography after the administration of intrathecal **cytarabine** ...	**cerebral vasospasm**	cytarabine	无关系	化学物质诱导疾病关系

作为自有知识。接着,采用推理建模、自适应阈值选取以及自有知识增强等一系列步骤,完成全文本级的关系抽取。实验结果表明,笔者所提出的模型在 CDR-revised 数据集上获得了优于现有大多数模型的性能。据笔者所知,本文是将自有知识融入到学术全文本关系抽取中的首次尝试,后续的研究计划主要有以下几个方面:①研究语言学上的逻辑推理,辅助关系推理,并进一步引入认知科学中的双系统理论提高学术全文本的关系抽取;②研究对比自有知识增强和外部知识增强在学术全文本关系抽取中的区别;③增加多个不同学科数据集下的模型性能对比分析。

参考文献:

[1] NUNOKAWA T, AKAZAWA M, YOKOGAWA N, et al. Late-onset scleroderma renal crisis induced by tacrolimus and prednisolone: a case report[J]. American journal of therapeutics, 2014, 21(5): e130－e133.

[2] ZHOU G D, SU J, ZHANG J, et al. Exploring various knowledge in relation extraction[C]//Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05). Michigan:ACL, 2005: 427－434.

[3] 李冬梅, 张扬, 李东远,等. 实体关系抽取方法研究综述[J]. 计算机研究与发展, 2020,57(7):25.

[4] 王嘉宁,何怡,朱仁煜,等. 基于远程监督的关系抽取技术[J]. 华东师范大学学报(自然科学版), 2020, 213(5):122－139.

[5] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre－training of deep bidirectional transformers for language understanding[EB/OL]. [2022－03－01]. https://arxiv. org/pdf/1810. 04805. pdf&usg = ALkJrhhzxlCL6yTht2BRmH9atgvKFxHsxQ.

[6] QUIRK C, POON H. Distant supervision for relation extraction beyond the sentence boundary[EB/OL]. [2022－03－01]. ht-

tps://arxiv. org/pdf/1609. 04873.
[7] PENG N, POON H, QUIRK C, et al. Cross-sentence n-ary relation extraction with graph lstms[J]. Transactions of the Association for Computational Linguistics, 2017, 5(1): 101 – 115.
[8] VERGA P, STRUBELL E, MCCALLUM A. Simultaneously self-attending to all mentions for full-abstract biological relation extraction[EB/OL]. [2022 – 01 – 09]. https://arxiv. org/pdf/1802. 10569.
[9] NAN G, GUO Z, SEKULIĆ I, et al. Reasoning with latent structure refinement for document – level relation extraction[EB/OL]. [2022 – 01 – 09]. https://arxiv. org/pdf/2005. 06312.
[10] LIU Y, OTT M, GOYAL N, et al. Roberta: a robustly optimized bert pretraining approach[EB/OL]. [2022 – 01 – 09]. https://arxiv. org/pdf/1907. 11692. pdf%5C.
[11] BELTAGY I, LO K, COHAN A. Scibert: a pretrained language model for scientific text[EB/OL]. [2022 – 01 – 10]. https://arxiv. org/pdf/1903. 10676.
[12] EVANS J S B T, FRANKISH K E. In two minds: dual processes and beyond[M]. Oxford: Oxford University Press, 2009.
[13] 薛露, 宋威. 基于动态标签的关系抽取方法[J]. 计算机应用, 2020, 40(6):1601 – 1606.
[14] 孙长志. 基于深度学习的联合实体关系抽取[D]. 上海:华东师范大学, 2020.
[15] LIN Y, SHEN S, LIU Z, et al. Neural relation extraction with selective attention over instances[C]//Proceedings of the 54th annual meeting of the Association for Computational Linguistics. Berlin: ACL,2016: 2124 – 2133.
[16] ZENG D, LIU K, LAI S, et al. Relation classification via convolutional deep neural network[C]//Proceedings of COLING 2014, the 25th international conference on computational linguistics. Dublin: Dublin City University and Association for Computational Linguistics,2014: 2335 – 2344.
[17] WANG L, CAO Z, DE MELO G, et al. Relation classification via multi-level attention cnns[C]//Proceedings of the 54th annual meeting of the Association for Computational Linguistics. Berlin: ACL, 2016: 1298 – 1307.
[18] SOARES L B, FITZGERALD N, LING J, et al. Matching the blanks: distributional similarity for relation learning[EB/OL]. [2022 – 03 – 10]. https://arxiv. org/pdf/1906. 03158.
[19] YAO Y, YE D, LI P, et al. DocRED: a large-scale document – level relation extraction dataset[EB/OL]. [2022 – 01 – 10]. https://arxiv. org/pdf/1906. 06127.
[20] GUPTA P, RAJARAM S, SCHÜTZE H, et al. Neural relation extraction within and across sentence boundaries[C]//Proceedings of the AAAI conference on artificial intelligence. Hawaii: AAAI, 2019, 33(1): 6513 – 6520.
[21] XU W, CHEN K, ZHAO T. Discriminative reasoning for document-level relation extraction[EB/OL]. [2022 – 01 – 10]. https://arxiv. org/pdf/2106. 01562.
[22] ZHOU H, XU Y, YAO W, et al. Global context-enhanced graph convolutional networks for document-level relation extraction[C]//Proceedings of the 28th international conference on computational linguistics. Barcelona: International Committee on Computational Linguistics, 2020: 5259 – 5270.
[23] CHRISTOPOULOU F, MIWA M, ANANIADOU S. Connecting the dots: document – level neural relation extraction with edge-oriented graphs[EB/OL]. [2022 – 01 – 10]. https://arxiv. org/pdf/1909. 00228.
[24] ZENG S, XU R, CHANG B, et al. Double graph based reasoning for document – level relation extraction[EB/OL]. [2022 – 01 – 10]. https://arxiv. org/pdf/2009. 13752.
[25] GUO Z, ZHANG Y, LU W. Attention guided graph convolutional networks for relation extraction[EB/OL]. [2022 – 01 – 10]. https://arxiv. org/pdf/1906. 07510.
[26] ZHOU W, HUANG K, MA T, et al. Document – level relation extraction with adaptive thresholding and localized context pooling[EB/OL]. [2022 – 03 – 01]. https://www. aaai. org/AAAI21Papers/AAAI-8308. ZhouW. pdf
[27] YANG A, WANG Q, LIU J, et al. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension[C]//Proceedings of the 57th annual meeting of the Association for Computational Linguistics. Florence: ACL, 2019: 2346 – 2357.
[28] WANG C, JIANG H. Explicit utilization of general knowledge in machine reading comprehension[EB/OL]. [2022 – 03 – 01]. https://arxiv. org/pdf/1809. 03449.
[29] 王冠颖. 融合知识图谱嵌入的关系抽取[D]. 杭州:浙江大学,2019.
[30] CHEN Q, ZHU X, LING Z H, et al. Neural natural language inference models enhanced with external knowledge[EB/OL]. [2022 – 03 – 01]. https://arxiv. org/pdf/1711. 04289.
[31] WANG Z, LI L, ZENG D, et al. Knowledge-enhanced natural language inference based on knowledge graphs[C]//Proceedings of the 28th international conference on computational linguistics. Barcelona: International Committee on Computational Linguistics, 2020: 6498 – 6508.
[32] PETERS M E, NEUMANN M, LOGAN IV R L, et al. Knowledge enhanced contextual word representations[EB/OL]. [2022 – 03 – 01]. https://arxiv. org/pdf/1909. 04164.
[33] YAO Y, YE D, LI P, et al. DocRED: A large-scale document – level relation extraction dataset[EB/OL]. [2022 – 03 – 01]. https://www. researchgate. net/profile/Zhenghao-Liu/publication/333815327_DocRED_A_Large-Scale_Document-Level_Relation_Extraction_Dataset/links/5fc60274299bf1a422c77e3d/DocRED-A-Large-Scale-Document-Level-Relation-Extraction-Dataset. pdf.
[34] BELTAGY I, LO K, COHAN A. SciBERT: A Pretrained language model for scientific text[C]//Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Hong Kong: ACL, 2019:3615 – 3620.
[35] XUE F, SUN A, ZHANG H, et al. GDPNet: refining latent multi-View graph for relation extraction[C]//Thirty-Fifth AAAI Confer-

ence on Artificial Intelligence. Online:AAAI,2021:2 - 9.
[36] LEE J, LEE I, KANG J. Self-attention graph pooling[C]//International conference on machine learning. Long Beach: ICML, 2019: 3734 - 3743.
[37] 白龙，靳小龙，席鹏弼，等. 基于远程监督的关系抽取研究综述[J]. 中文信息学报，2019，33(10)：10 - 17.
[38] 马进，杨一帆，陈文亮. 基于远程监督的人物属性抽取研究[J]. 中文信息学报，2020，34(6)：64 - 72.
[39] 谌予恒,王峥. 结合注意力机制与残差网络的远程监督关系抽取[J]. 计算机与数字工程,2020,48(4)：909 - 913.
[40] 高勇. 基于远程监督的企业实体关系抽取算法的研究与改进[D]. 上海:上海市计算技术研究所,2020.
[41] ZHANG N, CHENn X, XIE X, et al. Document-level relation extraction as semantic segmentation[EB/OL]. [2022 - 03 - 01]. https://arxiv. org/pdf/2106. 03618. pdf? ref = https://githubhelp. com.
[42] Paperswithcode. Relation extraction | papers with code[EB/OL]. [2022 - 01 - 14]. https://paperswithcode. com/task/relation-extraction#datasets.
[43] YAO Y, YE D, LI P, et al. DocRED: a large-scale document - level relation extraction dataset[EB/OL]. [2022 - 03 - 01]. https://aclanthology. org/P19-1074/? ref = https://githubhelp. com.
[44] LI J, SUN Y, JOHNSON R J, et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction[J]. Database, 2016(1):1 - 10.
[45] WU Y, LUO R, LEUNG H C M, et al. Renet: a deep learning approach for extracting gene-disease associations from literature[C]//International conference on research in computational molecular biology. Berlin: Springer, 2019: 272 - 284.
[46] MICIKEVICIUS P, NARANG S, ALBEN J, et al. Mixed precision training[EB/OL]. [2022 - 03 - 01]. https://arxiv. org/pdf/1710. 03740. pdf? ref = https://githubhelp. com.
[47] VERGA P, STRUBELL E, MCCALLUM A. Simultaneously self-attending to all mentions for full-abstract biological relation extraction[EB/OL]. [2022 - 03 - 01]. https://arxiv. org/pdf/1802. 10569.
[48] CHRISTOPOULOU F, MIWA M, ANANIADOU S. Connecting the dots: document-level neural relation extraction with edge-oriented graphs[EB/OL]. [2022 - 03 - 01]. https://aclanthology. org/D19 - 1498/.
[49] ZHANG Z, YU B, SHU X, et al. Document-level relation extraction with dual-tier heterogeneous graph[C]//Proceedings of the 28th international conference on computational linguistics. Barcelona: International Committee on Computational Linguistics, 2020: 1630 - 1641.
[50] WANG D, HU W, CAO E, et al. Global-to-local neural networks for document-level relation extraction[EB/OL]. [2022 - 03 - 01]. https://arxiv. org/pdf/2009. 10359.
[51] BELTAGY I, LO K, COHAN A. Scibert: a pretrained language model for scientific text[EB/OL]. [2022 - 03 - 01]. https://aclanthology. org/D19-1371. pdf.

作者贡献说明：
卓可秋:提出研究思路与方法、论文初稿撰写；
沈思:提出论文思路及修订；
王东波:论文修订与审阅。

Research on Relation Extraction of Academic Full-Text Based on Self-Owned Knowledge Enhancement

Zhuo Keqiu[1] Shen Si[2] Wang Dongbo[1]
[1] School of Information Management, Nanjing Agricultural University, Nanjing 210095
[2] School of Economics and Management, Nanjing University of Technology, Nanjing 210094

Abstract: [**Purpose/Significance**] Relation extraction under academic full-text is the key technology for the construction of academic full-text knowledge graph. The constructed academic knowledge graph can realize the structure and knowledge of documents, and improve the efficiency of researchers retrieving documents, analyzing documents and grasping scientific research trends, and cognitive reasoning through graphs contributes to implicit knowledge discovery. [**Method/Process**] Enhancing relation extraction through external knowledge has achieved results in many studies, but relation extraction for specific fields often lacked available external knowledge. The research in this paper found that the high-confidence knowledge in the full-text could also be used to assist the extraction of full-text relations. For this reason, based on the dual-system theory of cognitive processes (system 1 is intuitive cognition, system 2 is reasoning cognition), this paper designed a sentence-level model to acquire knowledge, and obtained high-confidence knowledge through remote supervision, and then high-confidence knowledge was integrated into the final classification layer of the text-level deep learning model. [**Result/Conclusion**] On the biomedical academic full-text data set (CDR-revised), the F1 is about 11.13% higher than the current state-of-the-art model.

Keywords: academic full-text relation extraction self-owned knowledge enhancement knowledge graph